Environmental Research 178 (2019) 108634

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Lessons learned from the application of machine learning to studies on plant response to radio-frequency

Malka N. Halgamuge[a,*], Devra Davis[b]

[a] Department of Electrical and Electronic Engineering, The University of Melbourne, Victoria, 3010, Australia
[b] Environmental Health Trust, Teton Village, WY, 83025, USA

ARTICLE INFO

Keywords:
Radio-frequency
Machine learning
Plant
RF-EMF

ABSTRACT

This paper applies Machine Learning (ML) algorithms to peer-reviewed publications in order to discern whether there are consistent biological impacts of exposure to non-thermal low power radio-frequency electromagnetic fields (RF-EMF). Expanding on previous analysis that identified sensitive plant species, we extracted data from 45 articles published between 1996 and 2016 that included 169 experimental case studies of plant response to RF-EMF. Raw-data from these case studies included six different attributes: frequency, specific absorption rate (SAR), power flux density, electric field strength, exposure time and plant type (species). This dataset has been tested with two different classification algorithms: k-Nearest Neighbor (*k*NN) and Random Forest (RF). The outputs are estimated using k-fold cross-validation method to identify and compare classifier mean accuracy and computation time. We also developed an optimization technique to distinguish the trade-off between prediction accuracy and computation time based on the classification algorithm. Our analysis illustrates *k*NN (91.17%) and RF (89.41%) perform similarly in terms of mean accuracy, nonetheless, *k*NN takes less computation time (3.38 s) to train a model compared to RF (248.12 s). Very strong correlations were observed between SAR and frequency, and SAR with power flux density and electric field strength. Despite the low sample size (169 reported experimental case studies), that limits statistical power, nevertheless, this analysis indicates that ML algorithms applied to bioelectromagnetics literature predict impacts of key plant health parameters from specific RF-EMF exposures. This paper addresses both questions of the methodological importance and relative value of different methods of ML and the specific finding of impacts of RF-EMF on specific measures of plant growth and health. Recognizing the importance of standardizing nomenclature for EMF-RF, we conclude that Machine Learning provides innovative and efficient RF-EMF exposure prediction tools, and we propose future applications in occupational and environmental epidemiology and public health.

1. Introduction

Accessing, collecting, storing and analyzing data is an increasing issue for organizations that must effectively handle expanding volumes and varieties of data (so-called Big Data). Big data, for example, climate data, health records, and business transactions are usually described by a high volume from multiple sources, compositions and structures, and high speed of new data creation (Dietrich, 2015). Thus, big data require specific methods and techniques for processing and analysis, not generally employed in many systems. In responses to these challenges, new methods are being developed for data science to yield new learning from the thorough analysis of big data, integrating methods and ideas from different disciplines including computer science, engineering, and statistics (Provost and Fawcett, 2013).

Machine learning (ML) is a use of artificial intelligence (AI) that gives frameworks the capacity to take in and enhance information without being expressly programmed automatically. This focuses on the advancement of computer programs that can access data and utilize it to "learn" for themselves. ML learn from past computations to provide reliable, repeatable decisions and results (VoPham et al., 2018; Patel et al., 2015; Kononenko, 2001). ML is a subset of AI that instructs a computer to obtain knowledge, by extracting information iteratively and discerning patterns in raw data (Goodfellow et al., 2016) so-called learning algorithm.

The maintenance of plant biodiversity remains a growing concern for developed and developing nations, as does the continued build-out of communications technology with its concurrent exposures to non-thermal weak radio-frequency electromagnetic field (RF-EMF). Limited

* Corresponding author. Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, Victoria, 3010, Australia.
E-mail address: malka.nisha@unimelb.edu.au (M.N. Halgamuge).

https://doi.org/10.1016/j.envres.2019.108634
Received 3 May 2019; Received in revised form 25 July 2019; Accepted 6 August 2019
Available online 16 August 2019
0013-9351/ © 2019 Elsevier Inc. All rights reserved.

literature exists on how and if RF-EMF from multiple sources, including mobile phones, other wireless transmitting devices (WTD) and base station radiation could affect plants and trees. While the development of mobile phone and other wireless technology has been exponential in the past few years, interest in evaluating the public health and environmental impacts of such exposures lags considerably behind this growth. Machine Learning tools have been developed to extract patterns from large data sets that may not otherwise be apparent. In conventional analyses of data, model or algorithms are applied to gather data to yield specific calculations. In contrast, Machine Learning provides an entirely novel approach to any data-rich field by allowing algorithms to emerging from an intense assessment of the data. Thus, algorithms can iteratively learn from data to reveal novel insights (Kononenko, 2001). The fields of Machine Learning and big data analytics can be used to extract valuable patterns from raw data. It also enhances the use of automated tools to assist in other health studies (in-vitro, in-vivo and epidemiology) and understand how environmental attributes may affect the final decision.

All living things contain the same basic building blocks of DNA, whether plant or animal. The plant kingdom offers an array of endpoints for evaluation that is often overlooked in the search for biologically essential outcomes relevant to exposures to radio-frequency radiation. While debate continues to surround the issue of impacts of this exposure on human health, a remarkably consistent and little-known literature is emerging regarding how this same radiation can affect several well-documented pathological indicators in plants. Thus, (Vian et al., 2016) found that RF elicited a range of responses at the cellular, molecular and whole plant scale, including disruption of numerous metabolic activities (reactive oxygen species metabolism, $\alpha-$ and $\beta-$ amylase, Krebs cycle, pentose phosphate pathway, chlorophyll content, terpene emission, etc.) as well as altered gene expression (calmodulin, calcium-dependent protein kinase, and proteinase inhibitor), and reduced growth. Previous investigations showed that maize, roselle, pea, fenugreek, duckweeds, tomato, onions and mung-bean plants are especially sensitive to RF-EMFs, with greater sensitivity at frequencies between (i) 800 and 1500 MHz ($p < 0.0001$), (ii) 1500 and 2400 MHz ($p < 0.0001$) and (iii) 3500 and 8000 MHz ($p = 0.0161$ (Halgamuge, 2016). We also found that the Random Forest (RF) classification algorithm would be the best among other classification algorithms (e.g. RF, Bayes Net, Nave Bayes, Decision Table, JRip, OneR, J48) (Halgamuge, 2017).

In light of current expansions of the technology, the question of plant impacts of RF-EMF is of critical importance for both agriculture and public health. Without the luxury of awaiting definitive analyses, policymakers face a daunting challenge in interpreting the rapidly changing information in bioelectromagnetics. In order to enhance our understanding of the impacts on plants from RF-EMF, this paper extends previous analyses and applies key concepts of machine learning to data sets. Then compare the Random Forest with the K-Nearest Neighbor algorithm developed to elucidate growing non-thermal levels of RF-EMF may impact specific endpoints of plant physiology. Data set has been extracted from 45 peer-reviewed scientific publications published between 1996 and 2016 with 169 experimental case studies conducted in the scientific literature. Machine Learning techniques are applied in order to discern whether there are common effects in plants in response to RF radiation (Halgamuge, 2017).

Machine learning techniques have been used in big data analysis; nonetheless, the challenge is to determine if the current data on non-thermal low power radio-frequency electromagnetic fields (RF-EMF) is sufficient to produce a prediction model with multiple variables. The raw-data grasps crucial information, such as patterns and trends in rates of growth, fertilization, or changes in pathology, which can be used to advance decision-making and optimize policies.

The Random Forest (RF) classification algorithm provides the best classification algorithms. This paper compares RF with Nearest Neighbor (kNN) algorithms in order to determine which of these two widely used approaches is more efficient and effective in extracting information from multiple studies. This paper additionally constructs an optimization technique to recognize the tradeoff between prediction accuracy and computation time in view of the application.

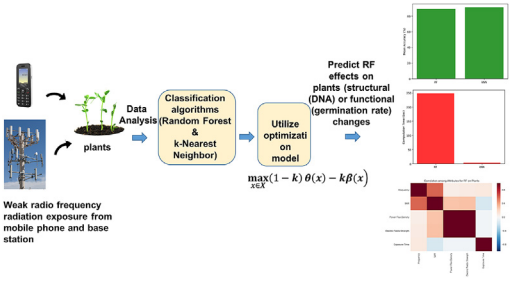
Weak radio frequency radiation exposure from mobile phone and base station
plants
Data Analysis
Classification algorithms (Random Forest & k-Nearest Neighbor)
Utilize optimizati on model
$\max_{x\in X}(1-k)\,\theta(x)-k\beta(x)$
Predict RF effects on plants (structural (DNA) or functional (germination rate) changes

Fig. 1. Graphical abstract.

The aim of this paper is twofold: to highlight established impacts of RF on plants across numerous studies and to compare the utility of two different methods of analyzing specific datasets on RF. Moreover, this paper developed an optimization technique to distinguish the trade-off between prediction accuracy and computation time based on the classification algorithm.

2. Materials and methods

For this study, we collected previous data (extracted data from 45 articles published between 1996 and 2016 that included 169 experimental case studies of plant response to RF-EMF) and used the Random Forest and the k-Nearest Neighbors classification algorithms for prediction. Fig. 1 shows the steps by step method to analyze the dataset, (classification algorithm and optimization model) to predict plant sensitivity to RF-EMF.

In our earlier research (Halgamuge, 2017), we indicated the Random Forest classification algorithm would be the best among several other classification algorithms (e.g. RF, Bayes Net, Nave Bayes, Decision Table, JRip, OneR, J48). Hence, this work compares the Random Forest algorithm with the k-Nearest Neighbor (kNN) algorithm using the same dataset.

2.1. The Best Accuracy-Computation-time pair (BAP)

Some applications, e.g. in medical fields, prediction of decision-making accuracy is essential than processing time. On the other hand, the time-sensitive fields require quick predictions such as natural disaster prediction. However, prediction accuracy percentage can also be used as a trade-off with processing time. Hence, we develop an optimization technique to distinguish the trade-off between prediction accuracy and computation time based on the classification algorithm.

Our purpose here is to acquire the BAP defined by the multiple objective unconstrained optimization problem given below

$$\max_{x\in \mathbf{X}}(1-k)\theta(x)-k\beta(x) \tag{1}$$

where x data split ratio (or k-fold number) between training and testing dataset where $x \in \mathbf{X}$, θ is the mean prediction accuracy or correctly classified instances and β is the computation time. Here $k \in [0,1]$, the weight factor shows the relative influence of the two objectives mean accuracy and computation time, which technique may be better depending on the application. For example, computational time matters as the size and complexity of the data increase in medicine and public health, accuracy is the key goal to manage sensitive data because a single prediction error can prompt a tragic result.

In order to solve the equation (1), our heuristic algorithm utilizes correct training and testing dataset to maximize prediction accuracy and minimize the computation time. The prediction accuracy is known

by the percentage of correctly classified instances completed in the shortest time.

2.2. Measurement

We evaluated the physiological or morphological effects of plants (bio-effects) or plant response (changed/unchanged or effect/no effect) due to exposure to weak radio-frequency radiation (SAR $4.9 \times 10^{-5} - 1.69$ W/kg and power flux density $0.011 - 0.332$ W/m^2) from mobile phones and base station defined as the changes in plant growth rate; (ii) seed germination rate (primary shoot and root length); (iii) thermographic imaging; (iv) carbohydrate metabolism; (v) oxidative damage/stress; (vi) gene expression; (vii) DNA damage; (viii) reactive oxygen species (ROS); (ix) cell function, enzyme activities; (x) mitotic index and mitotic abnormalities; (xi) mutation rates and genomic stability; (xii) pigmentation (chlorophyll concentration); and (xiii) chromosomal aberrations and micronuclei.

2.3. Data collection

We extracted data from 45 peer-reviewed scientific publications published between 1996 and 2016 that included 169 distinct experimental case studies to predict the potential responsiveness of plants to RF-EMF. The data employed in this analysis have been gathered from our previous review study (Tables 2–5 Halgamuge, 2016 (Halgamuge, 2016)) that extracted high levels of understanding from raw data using different classification algorithms and performance evaluation methods. The collected dataset comprises of 8 attributes of RF-EMF and 169 experimental case studies or instances such as frequency, specific absorption rate (SAR), power flux density, electric field strength, exposure time and plant type (species).

2.4. Data analysis

In our study, we utilize binary classification which classifies the data into two groups, e.g. whether or not the non-thermal low power radiofrequency electromagnetic fields (RF-EMF). Independent variables such as frequency, specific absorption rate (SAR), power flux density, electric field strength, exposure time and plant type (species) impact on sensitive plant species.

Data Inclusion criteria and data pre-processing criteria are as shown in our previous study (Halgamuge, 2017). The analysis is performed by using Python 3.6.0 on macOS Sierra (Version 10.12.6), 1.7 GHz Intel Core i7, 4 GB RAM, 1600 MHz DDR3.

2.4.1. Classification algorithms

This paper applies validated machine learning algorithms to determine whether there are consistent biological impacts of exposure to non-thermal low power radio-frequency electromagnetic fields (RF-EMF) utilizing previous experimental observations. Hence, it can be used to predict additional consequences. Here we assign to which class (C = effect, no-effect) a new observation belongs, utilizing two methods of machine learning algorithms: (i) k-Nearest Neighbor (kNN), and (ii) Random Forest classification algorithms. The capacity to precisely classify new observations is essential for reliable prediction. The distinctions between the two different machine learning algorithms that were applied to the assembled dataset are as follows: The k-Nearest Neighbor sorts information solely based on the class of the nearest neighbor. The Random Forest uses a decision tree framework that provides the automatic classification of data or new observation based on the number of trees in the forest.

K-Nearest Neighbors Algorithm: The kNN algorithm is one of the simplest machine learning algorithms which assumes that equal things exist in the closest neighbourhood. The kNN builds predictions from the training dataset, which is stored in the memory and then applied to new data sets. For example, to classify unknown information or data, kNN finds the set of data from the training information which is nearest to unknown information. The distance estimation is performed in kNN, which then assigns the highest voted classes out of these neighbouring classes. Determining the optimal estimate of k is essential, as it impacts the classifier performance. The best judgment of k depends upon the number of bytes of information; larger values of k decrease the impact of the noise on the classification (Everitt et al., 2011). The closeness of the new information with the training information can be estimated utilizing Euclidean distance (Everitt et al., 2011). The Euclidean distance d between two information sets focuses m and n are determined and integrated into the algorithm using $d(m, n) = \sqrt{\sum_{i=1}^{N} (m_i - n_i)^2}$.

Random Forest Algorithm: The Random Forest is a combination of tree-structured predictors that rely on tree classification algorithm. The key idea of driving the Random Forest algorithm is basic yet robust. The Random Forest comprises many decision trees that work as a group. Every single tree in the Random Forest predicts a class. A class (e.g. effect, no-effect) with the most votes becomes the foundation for predicting whether or not there are biological impacts from RF (Breiman, 2001).

The k-Nearest Neighbor sorts information solely based on the class of the nearest neighbor. Random Forest uses a decision tree framework that provides automatic classification of data.

2.4.2. K-fold cross-validation

The k-fold cross-validation method splits the data into ten equal parts and then uses the first nine parts for training and considers the final fold for testing purposes. The classification model performance uses a confusion matrix-10-folds cross-validation method (k = 10). This indicates through a comparative study between the classifiers which classifier is the best for the given dataset. Classification Accuracy is the accuracy of that technique as the number of right predictions on unseen data utilizing cross-validation from all predictions performed. It is generally the number of right predictions made divided by the aggregate number of predictions made, multiplied by 100 to transform it into a percentage.

2.4.3. Performance evaluation methods

Classification algorithms are utilized for training a data set to develop a model, and then the model can be used to designate unclassified instances into one of the distinct classes. A test set is used to decide the accuracy of the model. Usually, the presented dataset is split into the training data and tests data sets. The training set is used to establish the model, and test sets are then applied to verify it.

The data percentage split mode splits the dataset into training data testing it with different percentage ratios. In this test mode, the identified percentage of the training data: test data split ratio is given by 90%:10%, 80%:20%, 70%:30%, etc. Outputs are then associated with understanding the classifier performances utilizing: (i) percentages of correctly classified instances (PCC), and (ii) computational time or CPU time in seconds.

2.4.4. (i) Percentage of correct classifications (PCC)

The classification algorithms are usually estimated utilizing the percentage of correct classifications (PCC), and this is expressed as:

$$\gamma(i) = \begin{cases} 1 & \text{if } a_i = p_i \\ 0 & \text{else} \end{cases}$$

where a_i expresses the actual value and p_i expresses the predicted value for the i^{th} instance. Using (2.4.4), PCC is defined,

$$PCC = \sum_{i=1}^{n} \frac{\gamma_i}{n} \times 100 \tag{2}$$

2.4.5. (ii) Computation time (CPU time)

CPU time, the time taken to develop a model is estimated for every

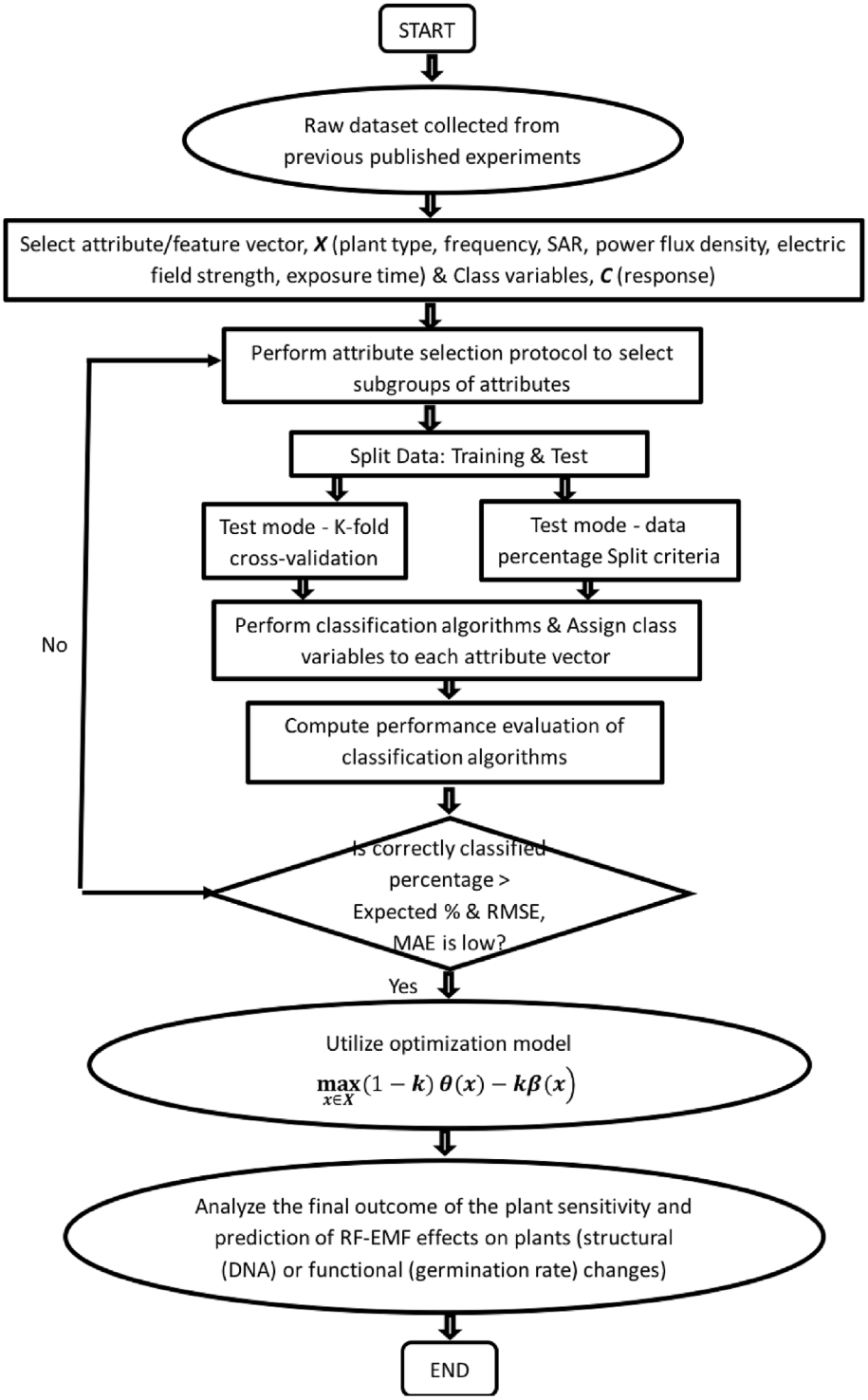
START
Raw dataset collected from previous published experiments
Select attribute/feature vector, X (plant type, frequency, SAR, power flux density, electric field strength, exposure time) & Class variables, C (response)
Perform attribute selection protocol to select subgroups of attributes
Split Data: Training & Test
Test mode - K-fold cross-validation
Test mode - data percentage Split criteria
Perform classification algorithms & Assign class variables to each attribute vector
Compute performance evaluation of classification algorithms
Is correctly classified percentage > Expected % & RMSE, MAE is low?
No
Yes
Utilize optimization model
$\max_{x \in X}(1-k)\,\theta(x) - k\beta(x)$
Analyze the final outcome of the plant sensitivity and prediction of RF-EMF effects on plants (structural (DNA) or functional (germination rate) changes)
END

Fig. 2. Flow chart to predict plant sensitivity to radio-frequency electromagnetic field (RF-EMF).

classification algorithm and is observed in the tool.

2.4.6. (i) Performance evaluation methods

Fig. 2 demonstrates the step by step method employed to conduct this analysis. This indicates how machine learning is applied to literature evaluating the impact of RF-EMF on various plant endpoints.

3. Results

Our previous work (Halgamuge, 2016) revealed that maize, roselle, pea, fenugreek, duckweeds, tomato, onions and mungbean plants are especially sensitive to RF-EMFs, with higher sensitivity at frequencies between (i) 800 and 1500 MHz ($p< 0.0001$), (ii) 1500 and 2400 MHz ($p< 0.0001$) and (iii) 3500 and 8000 MHz ($p= 0.0161$). Table 1 demonstrates the dataset that affected plants. However, we could not find a unified correlation on, such as which frequency influenced specific plants, as sometimes we observed the neglected impact of the same frequency.

Here we compare the Random Forest and k-Nearest Neighbor (*k*NN) algorithms based on the accuracy of the classifier and the computation

Table 1
Overview of the affected plants.

Plant	Scientific Name	Frequency (MHz)	Electric Field Strength (V/m)	Exposure Time (min)
Aspen	Populus	2000	–	80640
Brassicaceae	Arabidopsis thaliana	900, 1800	0.162	14400
Broadbean	Vicia faba L.	915	93.11, 115.1, 131.7	72
Celery	Apium graveolens	860	6.14	30240
Common wheat	*Triticum aestivum*	2450	6892.1	0.08
Dill	Anethum graveolens	860	6.14	30240
Duckweeds	Lemna minor	860, 900, 1287, 1900, 2000, 2500, 3000, 3500, 5500, 8000	1.7, 7.8, 10, 23, 41, 45, 50, 55, 60, 120	30, 60, 120, 240, 840, 1440
Fenugreek	Trigonella foenumgraecum	900, 1800	–	30, 60, 30, 60, 120, 240, 480
Flax	Linum usitatissimum L. var Ariane	900	61.4	120
Hyaciuth bean	Lablab, Fabaceae	1800	–	240, 1440
Lentil	Lens culinaris	1800	–	2880
Ligneous	Rosa hybrida	900	5, 200	30
Maize	Zea mays L.	900, 950, 1800	4.34, 11.187, 16.81, 19.41, 23.78	30, 60, 120, 240, 480, 720, 1440, 2160, 20160
Mologabean	Vigna, Papilionaceae	1800	–	240, 1440
Mungbean	Vigna radiata	900, 1800	0.056, 5.7	30, 60, 120, 240, 1440, 4320
Onions	Allium cepa bulbs	890, 900, 915	0.04, 0.43, 10, 23, 41, 120	120, 180, 240, 540
Parrotfeather	Myriophyllum aquaticum Verdc.	2000, 5500	15.65, 23, 30	60
Parsley	Petroselinum crispum	860	6.14	30240
Pea	Pisum sativum L.	900, 947.5, 1800	42.6	30, 60, 120, 240, 480
Plectranthus	Lamiaceae	902	42.5	60
Redbean	Vigna, Faboideae	1800	–	240, 1440
Riceplant	Oryza sativa L.	2280, 2490	8.5	60, 240, 420, 600
Roselle	Hibiscus Sabdariffa	2280, 2490	0.7, 0.8, 0.14, 0.47, 0.61, 0.72, 0.85, 0.9, 0.92, 1.05, 1.08 1.12, 8.5	43200
Soybean	Glycine max	900, 1800	0.56, 5.7, 41	120, 240, 1440
Tomato	Lycopersicon esculentum, VFN-8	900, 1250, 2450	0.6, 3.9, 5, 6, 689.2	0.013, 2, 10, 4320, 14400, 3153600

time. With the increase in computational time, we can now choose algorithms that are capable of conducting very intensive calculations. This study has classified RF on plant data using k-fold cross-validation test option to predict the expected outcomes by comparing Random Forest and k-Nearest Neighbor (*k*NN).

Figs. 3 and 4 show the comparison of each classifier output based on mean accuracy and computation times. Our analysis illustrates *k*NN classifiers (91.17%) and Random Forest (89.41%) perform relatively the same in terms of mean accuracy, nonetheless, the computation time to train the model was observed lower in *k*NN (3.38 s) than Random Forest (248.12 s). Computational time is significantly lower, utilizing *k*NN technique, indicating that *k*NN is better than Random Forest classifier, with a preferred Best Accuracy-Computation-time Pair (BAP).

There is a direct association between the number of trees in the

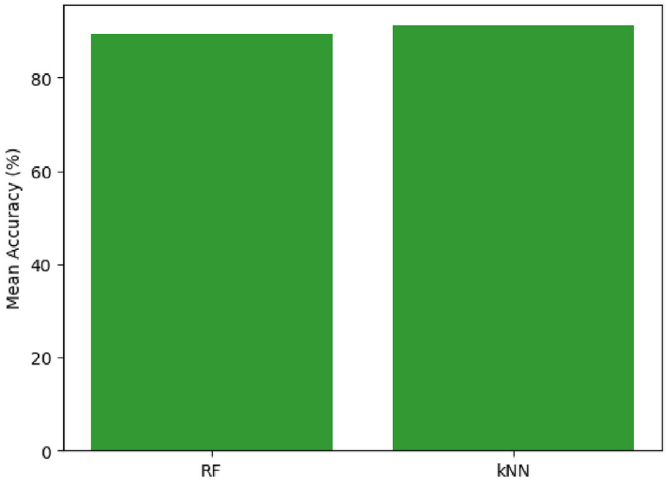

Fig. 3. Comparing prediction accuracy of Classifier (RF - Random Forest and *k*NN - *k*-Nearest Neighbor.

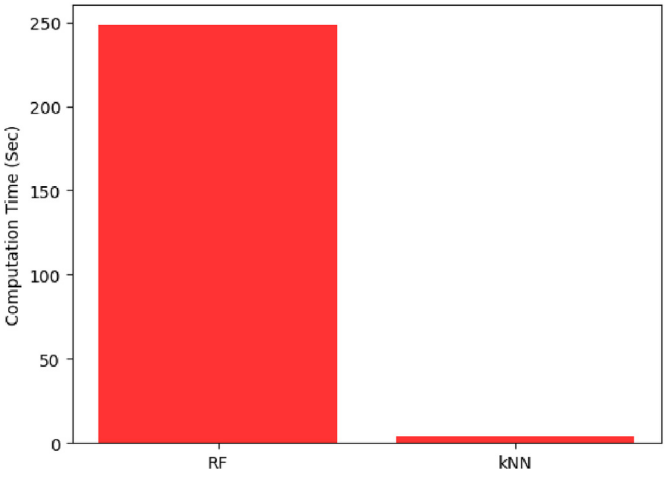

Fig. 4. Comparing computation time of Classifier (RF - Random Forest and *k*NN - *k*-Nearest Neighbor).

Table 2
Number of trees in the analysis of computation time and accuracy of the Random Forest classifier.

Trees/node	Mean Accuracy (%)	Computation Time (sec)
5	88.82	1.956
105	89.41	41.23
205	89.41	80.77
305	89.41	117.29
405	89.41	156.43
505	89.41	197.29
605	89.41	240.28
705	89.41	279.73
805	89.41	362.83
905	89.41	408.84

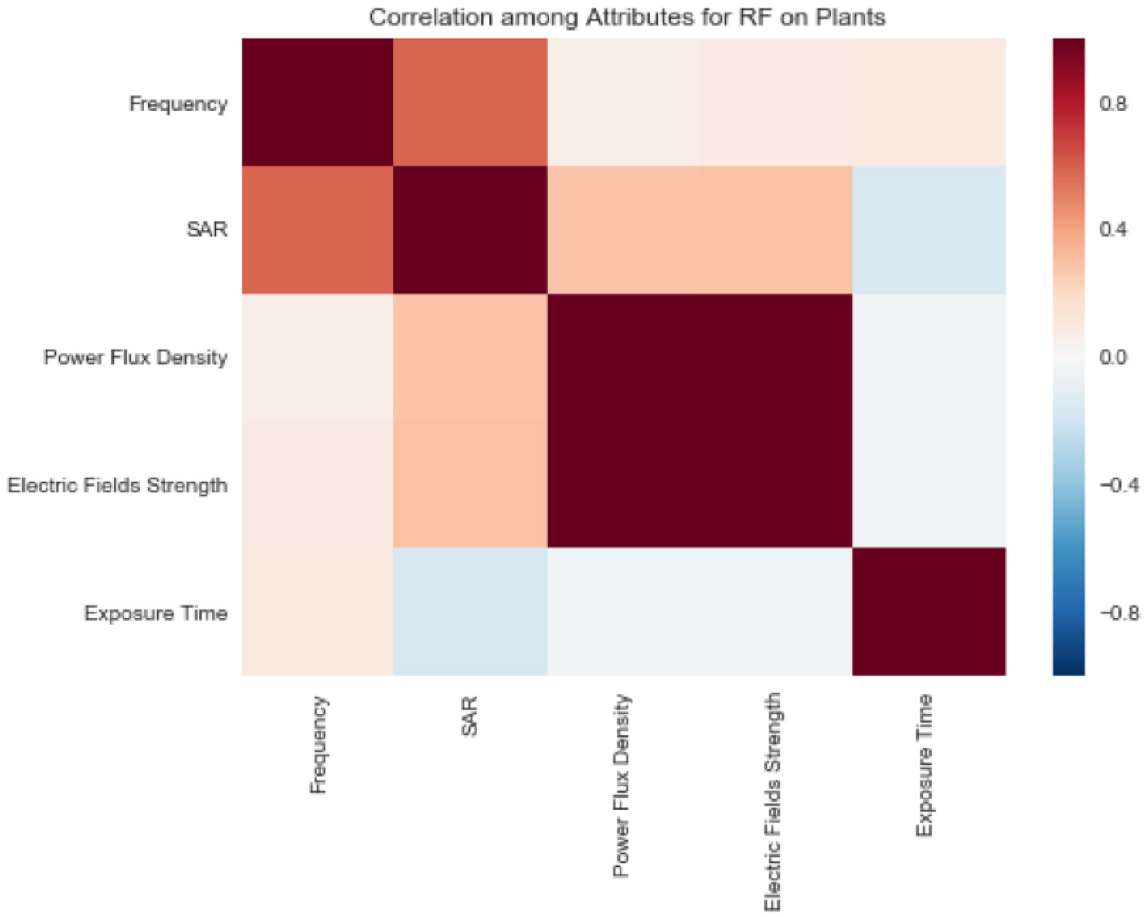

Fig. 5. Correlations among attributes for RF-EMF on plants, with red/pink indicating strong correlation and blue signaling no correlation. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

forest and the accuracy of the prediction obtain, the higher the number of trees, the outcome is more accurate (Everitt et al., 2011). Hence, we observe that correlation. Table 2 shows the effect of the number of trees on computation time and accuracy of Random Forest classifier. This reveals that expanding the number of trees does not change mean accuracy fundamentally. In contrast, an increased number of trees results in substantially higher computation time. Moreover, direct association between the number of trees in the forest, the accuracy of the prediction, and the amount of time required to calculate that prediction was observed. However, the accuracy of production is not improved once one goes above 160 studies.

Furthermore, as shown in Fig. 5, a very strong correlation was observed between SAR with frequency, as well as SAR with power flux density and electric field strength. The relationship between SAR and exposure time, however, was not significant.

4. Discussion

Extending on previous analysis that identified sensitive plant species, we extracted data from 45 articles published (1996–2016) that carried 169 experimental case studies of plant response to RF-EMF. Raw-data from these case studies included six different attributes: frequency, specific absorption rate (SAR), power flux density, electric field strength, exposure time and plant type (species). However, the response rate of the plant for RF depends on many other factors, such as the phenological stage of the plant, the geographical region of the plant growth.

This paper gives a comparative investigation of classification algorithms on plant data sets. These contain Random Forest and k-Nearest Neighbors classification algorithms and estimated their efficiency and computation time. The importance of the Environmental Epidemiology is to examine the health consequences related to environmental exposures and apply existing knowledge to decrease the exposure of specific populations to known hazard factors. Furthermore, this paper tries to translate the principles and methods of machine learning to the world of occupational and environmental epidemiology, and public health.

In this brief paper, we have applied two algorithms of machine learning to data extracted from 45 peer-reviewed scientific publications published between 1996 and 2016 with 169 experimental case studies in order to predict patterns and trends of plant responsiveness on non-thermal weak radio-frequency electromagnetic fields (RF-EMF). Although, our previous work was conducted only for in-vitro studies, in the future, as more studies become available. Researchers could apply machine learning algorithms to long-term laboratory experiments using whole organisms (in-vivo), and epidemiology studies in order to enhance the accuracy of prediction (Halgamuge, 2017). Further research is vital in order to understand whether and how some RF-EMF attributes (e.g. frequency, SAR, exposure time, power flux density) affect the prediction of responses in living beings. The limitation of this investigation is that the relatively low sample size (169 reported experimental case studies) restricts the strength of conclusions. In our previous study (Halgamuge, 2017) more Machine Learning algorithms (Bayes Net, NaiveBayes, Decision Table, JRip, OneR, J48, Random Tree, and Random Forest) have been examined, however, observed insignificant performance contrast with Random Forest Algorithm. Early investigations (Halgamuge et al., 2015; Gustavino et al., 2016; Rammal et al., 2014) demonstrated that the rate of germination was

substantially reduced in soybean, broadbean, tomato plant species when exposed to 0.56. 5.7, 41, 93, 6 V/m power and 900, 915, 1250 MHz frequencies exposure over 3–10 days.

The execution of the Random Forest algorithm increments expands with the number of trees and converges after some point. An extensive number of trees decreases the risk of overfitting and variance in the model. It should be noted that at some point, the abundance of trees makes model training wasteful (Patel et al., 2015) increasing computation time, which results in far more substantial costs of execution. Our examination does not cover memory utilization for this situation. Nonetheless, it can be noted that the substantial number of trees consumes a larger RAM space (Patel et al., 2015) when we use the Random Forest technique.

A key limitation with bioelectromagnetics research is that electric and magnetic field exposure conditions and nomenclature are not sufficiently provided or standardized with adequate accuracy to empower reproducible outcomes (Dlugosz, 2013). In vitro findings are essential to investigate natural and induced occurrences, however, the energies and induced impacts results from confounding factors are hard to avoid. For instance, non-homogenous background electromagnetic field situations and temperatures inside laboratory incubators have been exhibited to skew outcomes (Portelli et al., 2013), and this essential critique can sensibly be applied to numerous studies that show or do not demonstrate biological effects. On the other hand, organisms have systems for rapidly repairing damage and looking after homeostasis (Barnes and Greenebaum, 2016). These natural systems can possibly make it harder to detect varying variable responses (McKee, 2018) and measure the subtle effects of weak externally applied fields. Choosing a classifier depends on the necessities of the application. Expanding the number of features attempted by classifier likewise builds the component space dimension causing “curse of dimensionality” which can reduce accuracy and increase computation time (Singh et al., 2017). Also, classifier performance relies upon attributes of classified data sets. Different correlations have been made on various classifier performance over different datasets to discover reasonable classifier for a given application. Certainly, even with high performing computers taking care of complex issues requires most appropriate classification techniques to reduce wastage of time and resources (Singh et al., 2017).

Regression modeling is a standout amongst the most imperative statistical techniques utilized as a part of analytical epidemiology and many other disciplines of public health. However, for evaluating large data sets research has indicated (Couronne, 2017) the Random Forest performed better than logistic regression method according to the considered accuracy. This technique may be better applied to studies of genetic and other biomarkers of exposure or early onset of disease as they are frequently utilized when endpoints are far away temporally or causally.

Machine learning can provide a remarkable tool since it checks some unexplained correlations in various attributes in any application (Halgamuge, 2017; Singh et al., 2017; Gupta et al., 2016). These techniques appear to be useful tools for identifying key parameters and impacts of RF-EMF on plants. Nonetheless, the accuracy of the estimation achieved gives a clear picture of algorithm performance over various data type and feature dimensions. Prediction in health division requires a more noticeable level of accuracy for enhanced diagnosis and treatment, while disaster management requires less computation time in prediction to take activities timely saving lives.

Future work for Bioelectromagnetics could apply algorithms to data sets in which appropriate variables are taken with an identical methodology (e.g. field strength, polarity, SAR, radiation frequency, damage types found, species affected, distance to radiation source etc). However, records of more individual experiments are needed to improve the sample size and predictive power of results (Halgamuge, 2017). As with all analyses the quality of the predictions that are made depends entirely on the quality of the dataset. The results obtained by our study show negligible (0.02% fluctuation among correctly classified categories, indicating that the results are significant. Even though each single reported experimental case study carried out between 1996 and 2016 (sample size = 169), may not be highly significant in a statistical sense, and they nonetheless are significant statistically when combined in the current framework. Thus, this analysis further provides insight into the potential utility of Machine Learning in Bioelectromagnetics. Further evaluations should be carried out as larger sample sizes become available over time.

4.1. Future direction

New approaches are combining data on direct or indirect exposures (such as sales of inhalers as an indicator of air pollution impacts) for investigating their relationship to various population health measures that have previously been hard to quantify. Significantly, georeferenced data are ending up more broadly accessible in low- and middle-income countries, diminishing obstructions for directing studies of occupational and environmental epidemiology, and public health in these countries. Such data incorporate those gathered through remote sensing, sensor networks, smartphones, and also the “internet of things” (network-connected smart objects) (Tonne et al., 2017). Modern technologies, (e.g. medical image, video, and sensors) also yield immense amounts of data on a single participant. The further growth of Big Data studies will rely on parallel processing on computer clusters or a cloud with corresponding platforms. In many instances, raw data will be discarded due to the difficulty of data storage and only relevant summaries will be stored. Big data created for purposes other than the research, of course, will also introduce challenges as far as data quality and representativeness of the population.

Further, advances in Artificial Intelligence (AI) facilitate, more accurate, high-resolution exposure modeling for environmental epidemiological studies, particularly with respect to high-performance computing to manage big data (big in space and time; spatiotemporal). In addition, further developments in Machine Learning and Deep Learning will ultimately be able to predict the measure of an environmental factor at a specific time and location (VoPham et al., 2018). Artificial Intelligence applications for environmental epidemiology can also be linked with other appropriate data as to outcomes, confounders, and so on., to explore regardless of whether specific environmental conditions related with a specific outcome for epidemiological research (VoPham et al., 2018). In the next two decades, transformative changes, specifically, the quick pace of technological advancement and data accessibility, will require public health researchers to make innovative use of existing large-scale data sets of health information and environmental exposures and their surrogates in order to enhance population health (Tonne et al., 2017). Possible future Geospatial artificial intelligence applications for occupational and environmental epidemiology and public health will necessarily employ cross-disciplinary approaches that scrub through different sources of high-quality Big Data (VoPham et al., 2018).

Finally, as a developing field, applying Artificial Intelligence, Machine Learning, Deep Learning and to occupational and environmental epidemiology, and population or public health requires interdisciplinary skill, including computer science, epidemiology, engineering, including biostatistics. This will require the build-up best practices for how to approach environmental exposure in light of the different biological, physical and chemical properties. More algorithms should be produced and utilized, and complex environmental data should be evaluated in order to discern changing patterns, structures, and relationships among key variables that may not be detectable through conventional analyses at this time.

We show the potential adaptability of Machine Learning in the field of Bioelectromagnetics research for plants and raise the prospects that ML could be applied to other topics such as in-vivo, In-vitro, and Epidemiological studies using animal, plant and human population.

5. Conclusion

Machine learning centers around the advancement of computer programs that can access data and utilize it to learn for themselves. They gain from past computations to create reliable, repeatable decisions and results. In this study, two main classifiers have been used to generate accurate prediction patterns and trends of plant responsiveness on non-thermal weak radio-frequency electromagnetic fields (RF-EMF) using previously published experimental data. Six different attributes depicted in this data set have been used for the analysis, including frequency, specific absorption rate (SAR), power flux density, electric field strength, exposure time and plant type (species). This study established an optimization strategy to recognize the trade-off between prediction accuracy and computation time in view of the application. Results indicate that *k*NN (91.17%) and RF (89.41%) perform relatively the same in terms of mean accuracy, nonetheless, the computation time to train the model was considerably reduced in *k*NN (3.38 s) than the Random Forest (248.12 s), which had a BAP nearly 100 times greater. This demonstrates that the *k*-Nearest Neighbor (*k*NN) classifier was preferable than Random Forest classifier for the plant data that we have collected. Hence, *k*NN can be progressively utilized to evaluate bioelectromagnetic studies to anticipate the sensitivity (bioeffects) of exposure to weak radiofrequency radiation impact on plants. Further advances should be possible through evaluating parameterized data sets that are part of larger sample sizes. Then Machine Learning is applied to a relatively small sample size of detailed studies of low statistical significance, this can and does yield helpful insights regarding consistently predicted biological responses from plant materials in response to bioelectromagnetic exposures. If sufficient numbers of individual studies can be aggregated, such as with recent reports on reactions of insect life to EMF and applications to environmental epidemiology, this may also yield useful results.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2019.108634.

References

Barnes, F.S., Greenebaum, B., 2016. Some effects of weak magnetic fields on biological systems: Rf fields can change radical concentrations and cancer cell growth rates. IEEE Power Electron 3 (1).
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1).
Couronne, R., 2017. Random Forest versus Logistic Regression: a Large-Scale Benchmark Experiment, Technical Report Number 205. University of Munich, Germany.
Dietrich, D., 2015. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. John Wiley & Sons, Indianapolis, IN, USA.
Dlugosz, T., 2013. Uncertainty analysis of selected sources of errors in bioelectromagnetic investigations. In: Proc. Of the 2nd International Conference on Biomedical Engineering and Biotechnology, Wuhan, China.
Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Miscellaneous clustering methods. In: Cluster Analysis, fifth ed. John Wiley & Sons, Chichester, UK.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. The MIT Press, Cambridge, MA. USA.
Gupta, A., Mohammad, A., Syed, A., Halgamuge, M.N., 2016. A comparative study of classification algorithms using data mining: crime and accidents in denver city the USA. Int. J. Adv. Comput. Sci. Appl. 7 (7), 374–381.
Gustavino, B., Carboni, G., Petrillo, R., Paoluzzi, G., Santovetti, E., Rizzoni, M., 2016. Exposure to 915 MHz radiation induces micronuclei in Vicia faba root tips. Mutagenesis 31 (2), 187–192.
Halgamuge, M.N., 2016. Review: weak radiofrequency radiation exposure from mobile phone radiation on plants. Electromagn. Biol. Med. 26 (2), 213–235.
Halgamuge, M.N., 2017. Machine learning for bioelectromagnetics: prediction model for weak radiofrequency radiation effect from mobile phone on plants. Int. J. Adv. Comput. Sci. Appl. 8 (11), 223–235.
Halgamuge, M.N., Yak, S.K., Eberhardt, J.L., 2015. Reduced growth of soybean seedlings after exposure to weak microwave radiation from GSM 900 mobile phone and base station. Bioelectromagnetics 36 (2), 87–95.
Kononenko, I., 2001. Machine learning for medical diagnosis: history, state of the art and perspective. Artif. Intell. Med. 23 (1), 89–109.
McKee, L., 2018. Meeting the imperative to accelerate environmental bioelectromagnetics research. Environ. Res. 23 (164), 100–108.
Patel, J., Shah, S., Thakkar, P., Kotecha, K., 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Syst. Appl. 42, 259–268.
Portelli, L.A., Schomay, T.E., Barnes, F.S., 2013. Inhomogeneous background magnetic field in biological incubators is a potential confounder for experimental variability and reproducibility. Bioelectromagnetics 34 (5), 337–348.
Provost, F., Fawcett, T., 2013. Data science and its relationship to big data and data-driven decision making. Big Data 1 (1), 51–59.
Rammal, M., Jebai, F., Rammal, H., Joumaa, W.H., 2014. Effects of long term exposure to RF/MW radiations on the expression of mRNA of stress proteins in lycopersicon esculentum. Transactions on Biology and Biomedicine 11 (10), 10–14.
Singh, A., Halgamuge, M.N., Lakshmiganthan, R., 2017. Impact of different data types on classifier performance of random forest, naïve bayes, and k-nearest neighbors algorithms. Int. J. Adv. Comput. Sci. Appl. 8 (12), 1–10.
Tonne, C., Basagaña, X., Chaix, B., Huynen, M., Hystad, P., Nawrot, T., Slama, R., Vermeulen, R., Weuve, J., Nieuwenhuijsen, M., 2017. New frontiers for environmental epidemiology in a changing world. Environ. Int. 104, 155–162.
Vian, A., Davies, E., Gendraud, M., Bonnet, P., 2016. Plant responses to high frequency electromagnetic fields. BioMed Res. Int. https://doi.org/10.1155/2016/1830262.
VoPham, T., Hart, J.E., Laden, F., Chiang, Y.Y., 2018. Emerging trends in geospatial artificial intelligence (geoai): potential applications for environmental epidemiology. Environ. Health 17, 40.